Heuristics for Relevancy Ranking of Earth Dataset Search Results

American Geophysical Union 2016

Christopher Lynnes (NASA ESDIS)
Patrick Quinn (Element 84)
James Norton (Element 84)

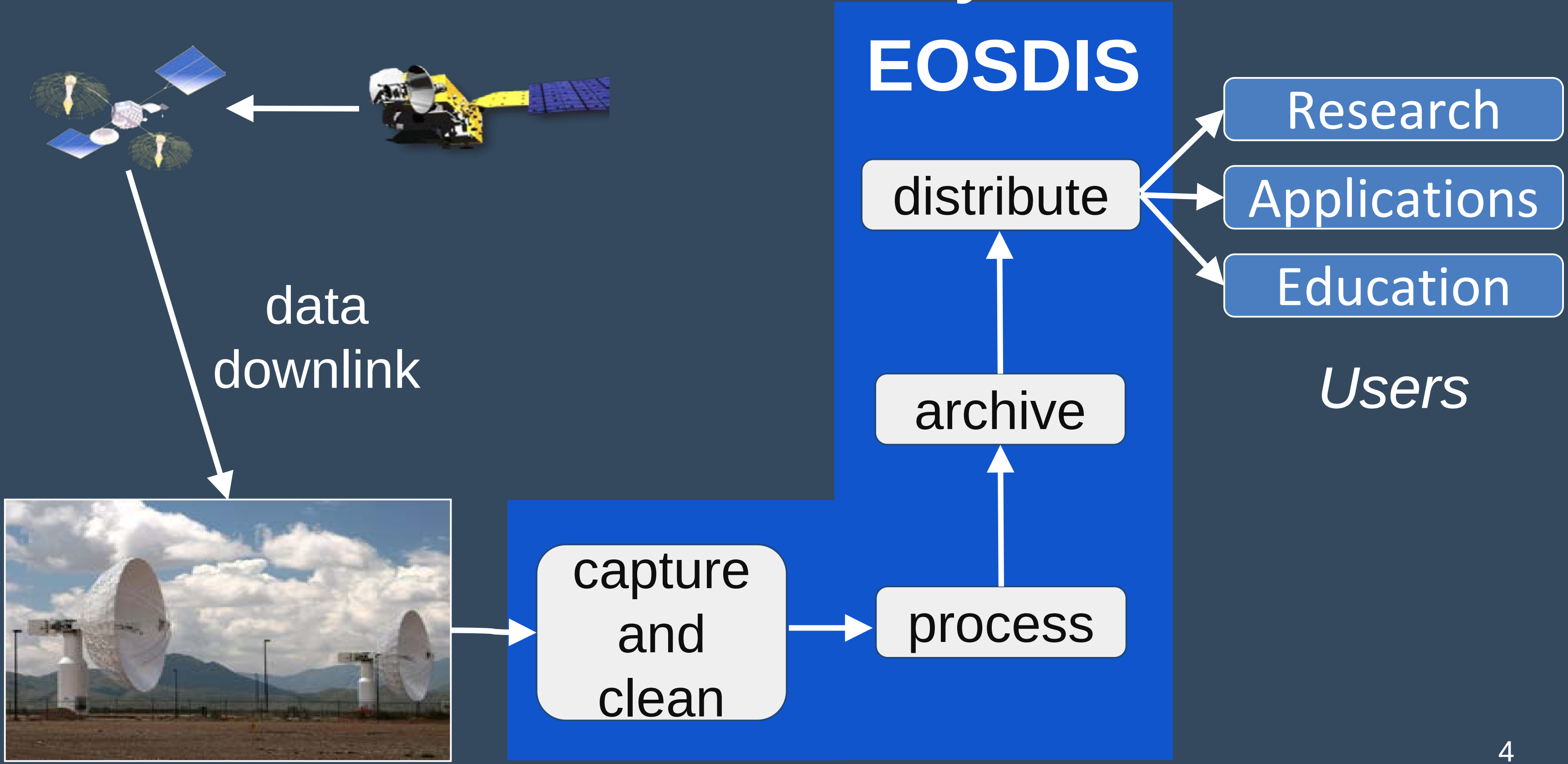
The Variety problem in Big Data from Satellites

Variety = Choice

Choice = Good

(Right?)

Earth Observing System Data and Information System



data downlink

EOSDIS

capture and clean

process

archive

distribute

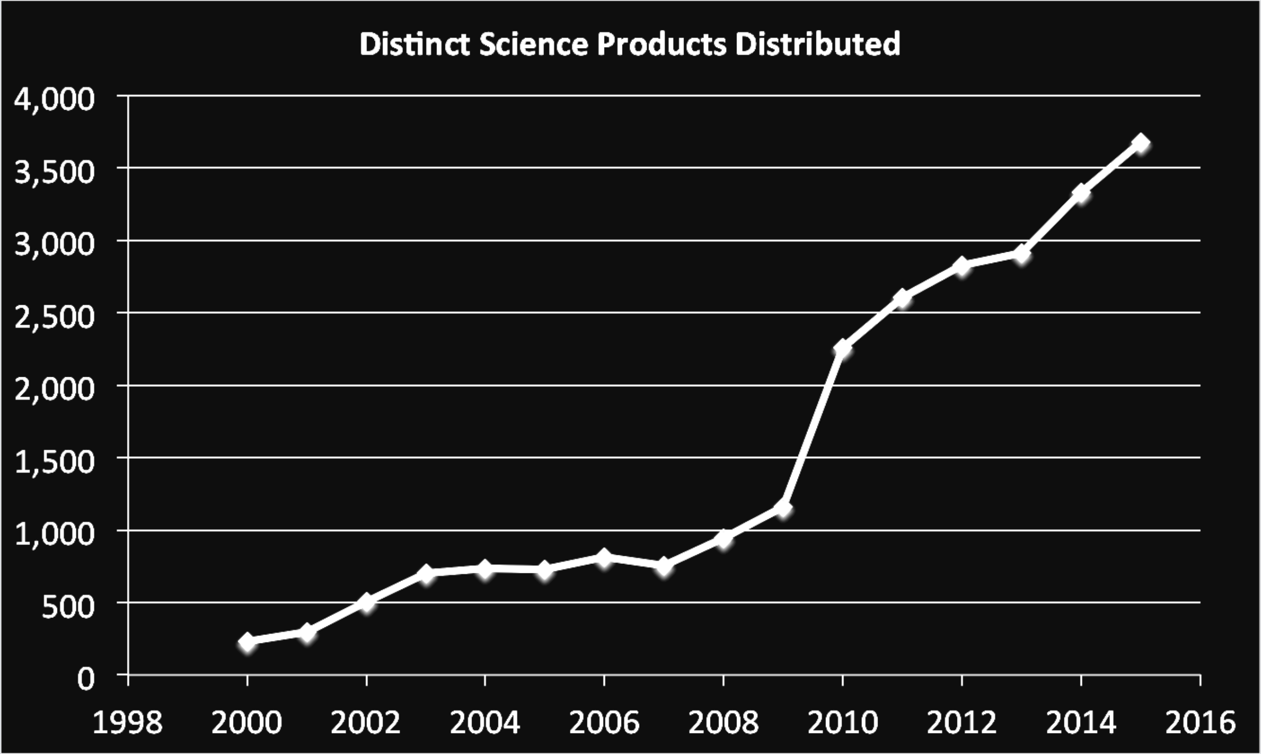
Research

Applications

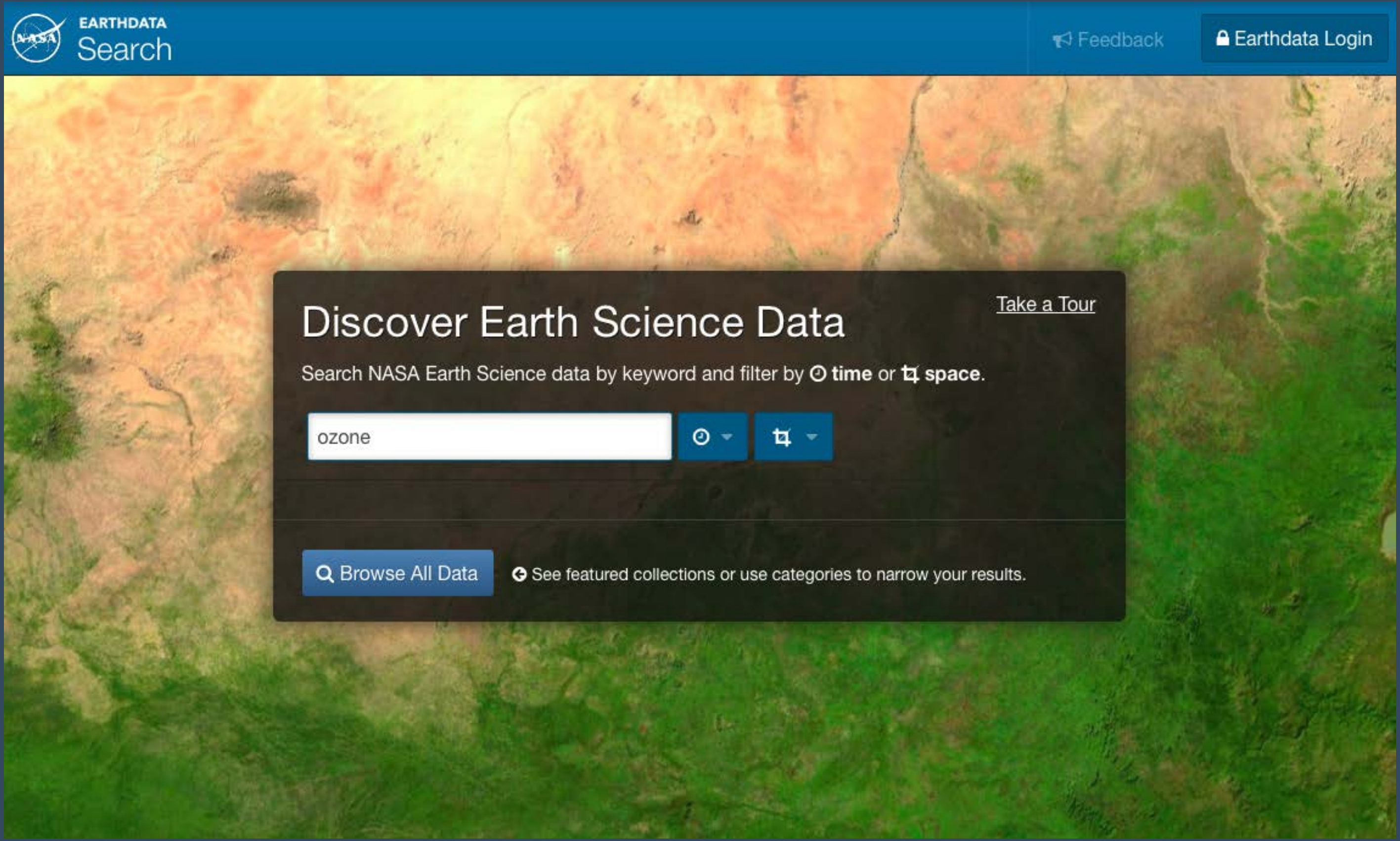
Education

Users

The Variety problem in Big Earth Data from Satellites



Earthdata Search Tool



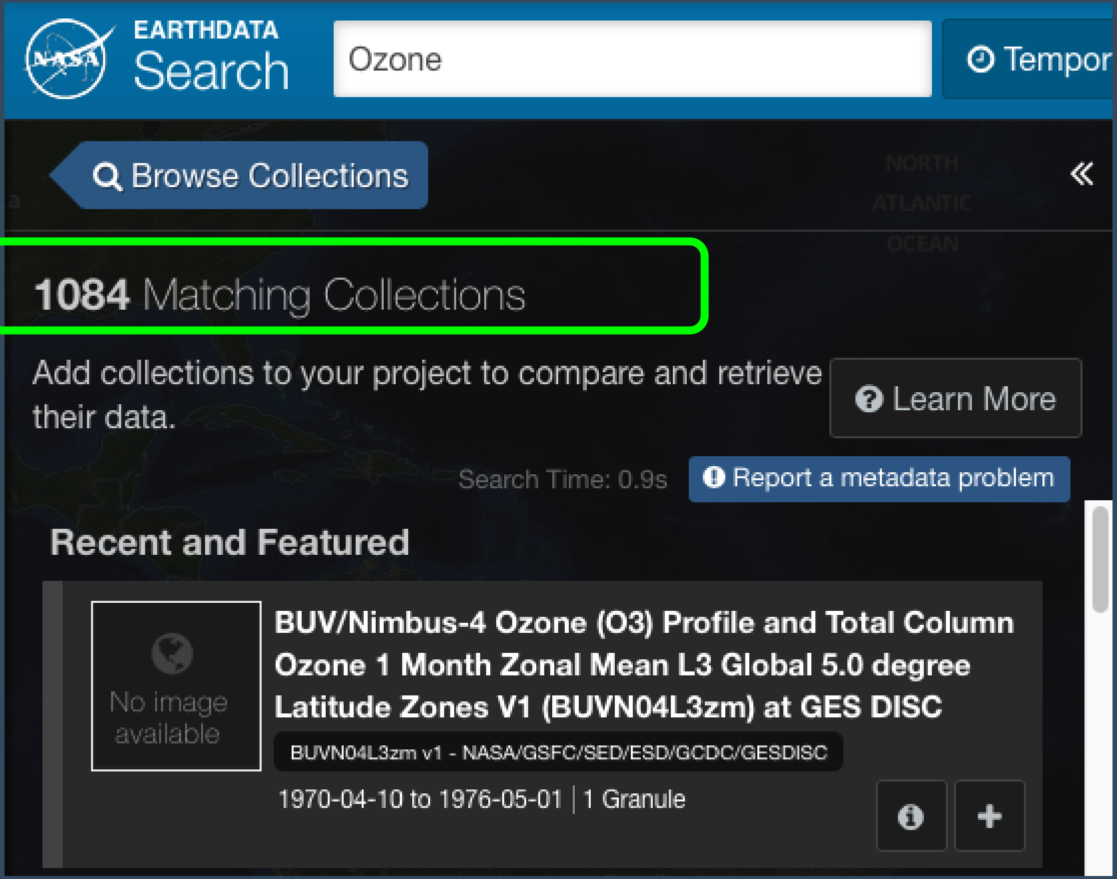
Discover Earth Science Data

Search NASA Earth Science data by keyword and filter by **time** or **space**.

ozone

See featured collections or use categories to narrow your results.

Too Many Datasets to Sift Manually



1084 Matching Collections

Add collections to your project to compare and retrieve their data.

Search Time: 0.9s

Recent and Featured

BUV/Nimbus-4 Ozone (O3) Profile and Total Column Ozone 1 Month Zonal Mean L3 Global 5.0 degree Latitude Zones V1 (BUVN04L3zm) at GES DISC

BUVN04L3zm v1 - NASA/GSFC/SED/ESD/GCDC/GESDISC

1970-04-10 to 1976-05-01 | 1 Granule

Where Does Variety Come From?

Instruments

- Fundamental differences: sounders, limb sounders, imagers...
- Incremental evolution in instrument design

Satellites: “Same” instrument on different satellites

Processing Level: Calibrated -> Swath -> Grid -> Model

Processing Algorithm

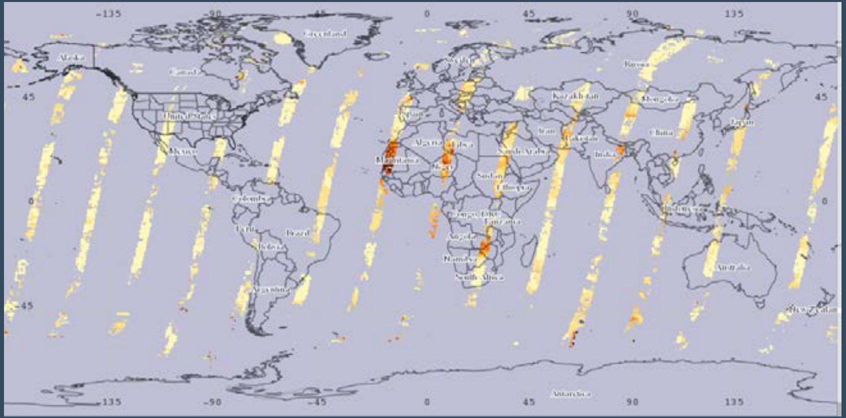
- Different basic principles
- Incremental evolution in algorithm development

Temporal Resolution: daily, multi-day, monthly, yearly

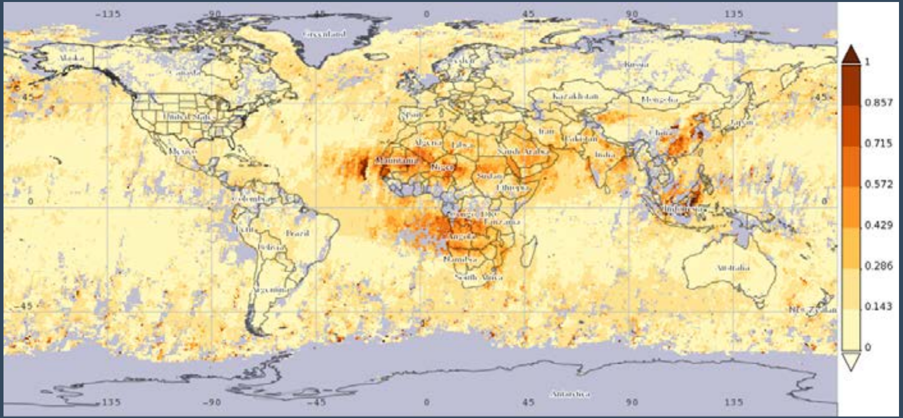
Spatial Resolution

Example: Time Aggregation

Aerosol Optical Depth at 555 nm from Multi-angle Imaging Spectro-Radiometer



Daily



Monthly

What To Do?

Emulate the best search engines:
return the most relevant results at the top of the list

Relevancy à la Wikipedia

“how well a retrieved document or set of documents meets the *information need of the user*”

HOW?

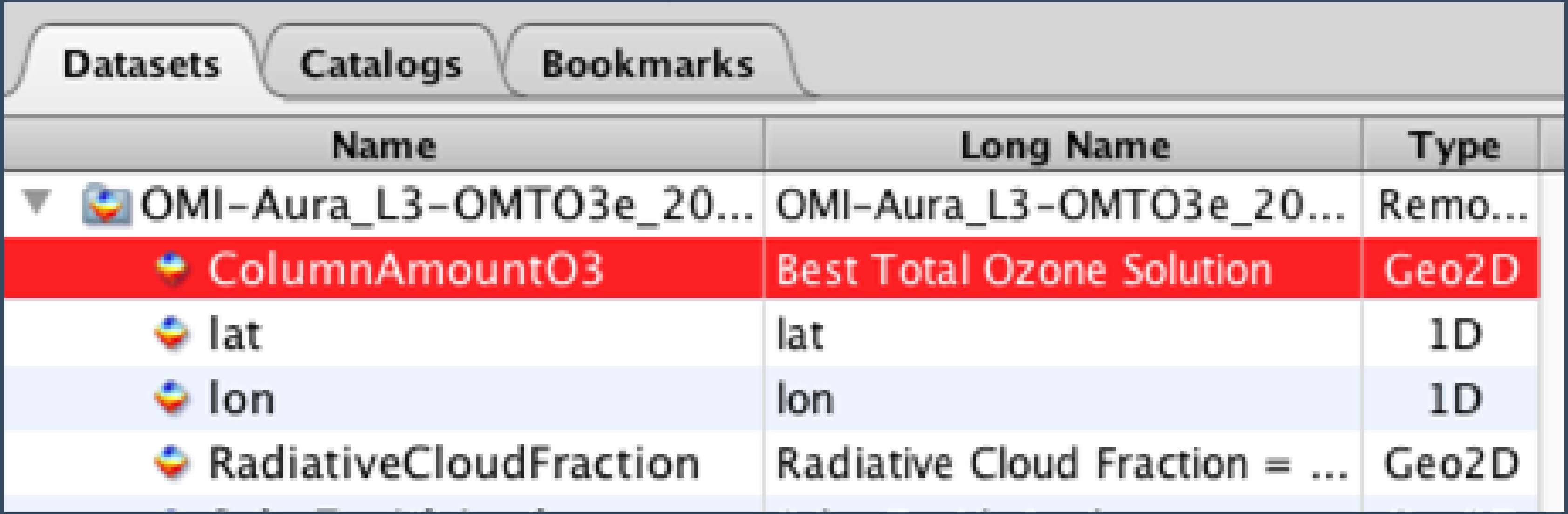
Relevancy Ranking Heuristics

Heuristic = “rule of thumb”

Basis is a quarter century of serving satellite data to researchers

The Content Heuristic*

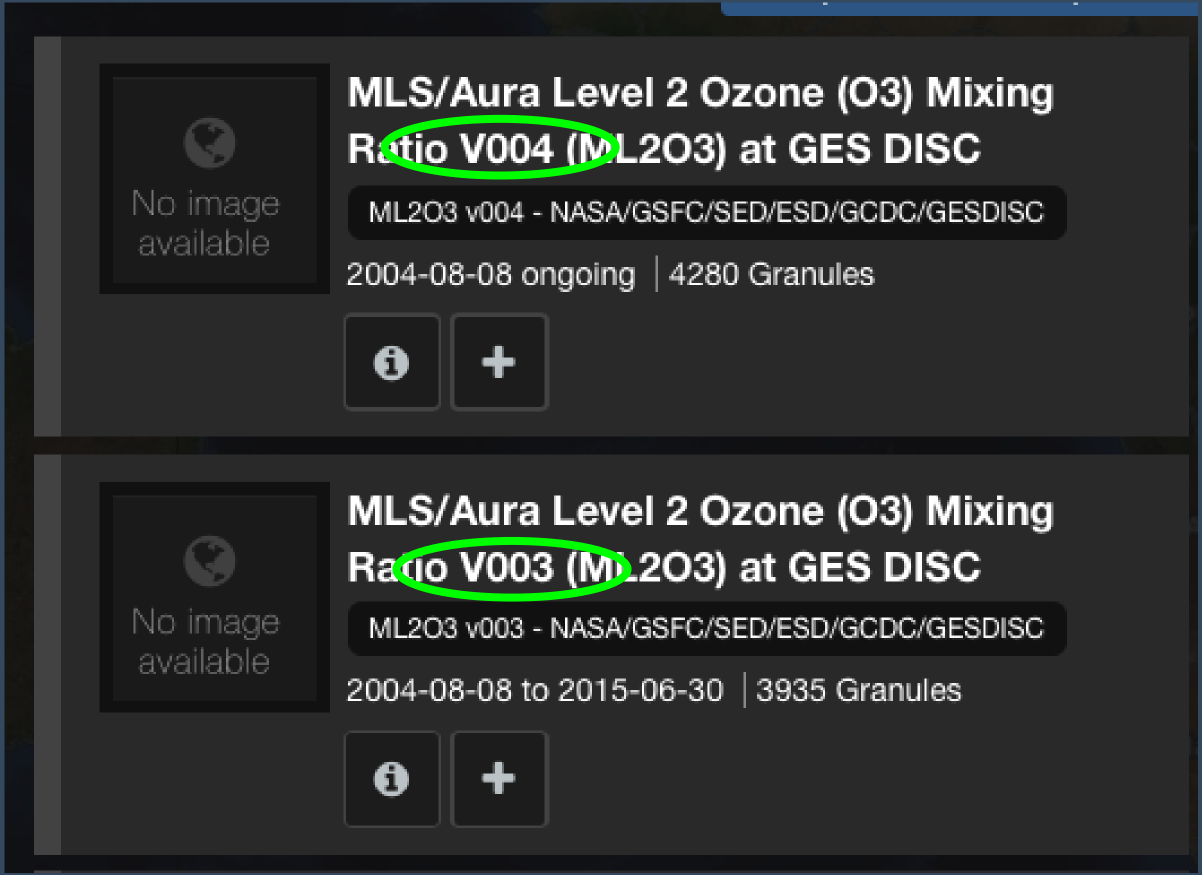
Got ozone?



Name	Long Name	Type
▼ OMI-Aura_L3-OMTO3e_20...	OMI-Aura_L3-OMTO3e_20...	Remo...
ColumnAmountO3	Best Total Ozone Solution	Geo2D
lat	lat	1D
lon	lon	1D
RadiativeCloudFraction	Radiative Cloud Fraction = ...	Geo2D

“New-and-improved” Heuristics

New-and-Improved Processing Version



MLS/Aura Level 2 Ozone (O3) Mixing Ratio V004 (ML2O3) at GES DISC

ML2O3 v004 - NASA/GSFC/SED/ESD/GCDC/GESDISC

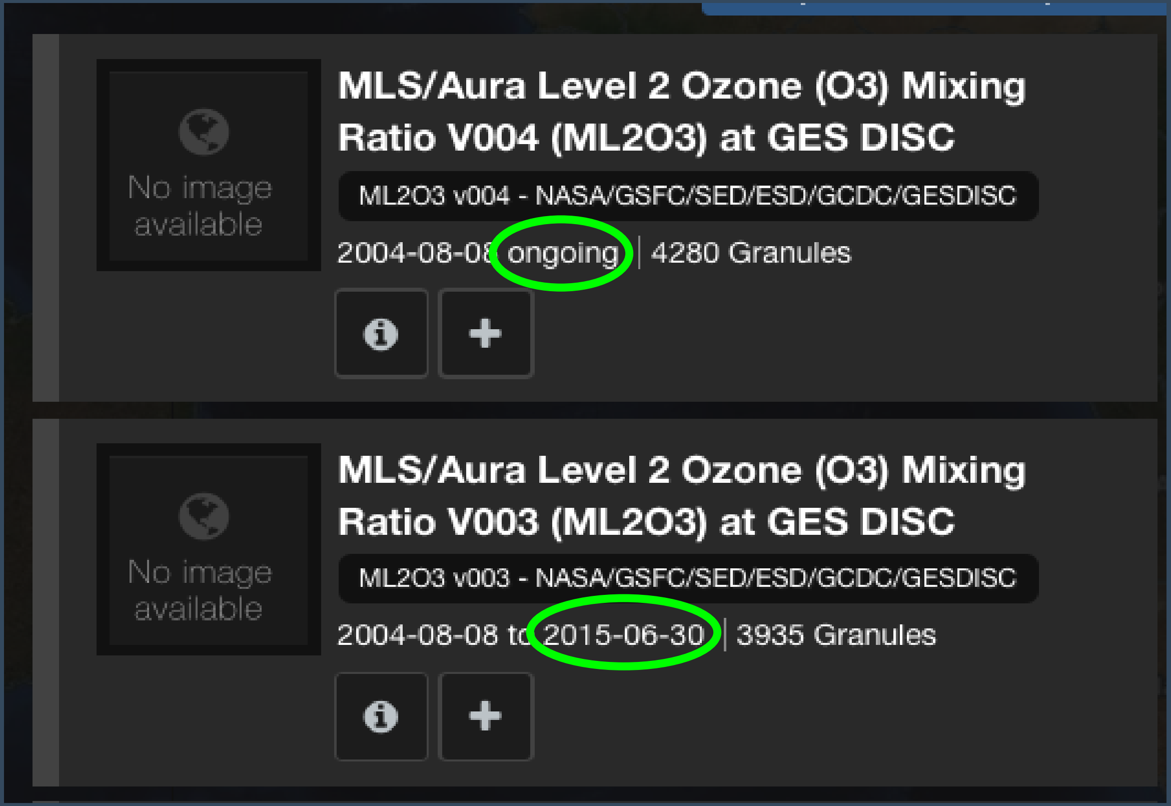
2004-08-08 ongoing | 4280 Granules

MLS/Aura Level 2 Ozone (O3) Mixing Ratio V003 (ML2O3) at GES DISC

ML2O3 v003 - NASA/GSFC/SED/ESD/GCDC/GESDISC

2004-08-08 to 2015-06-30 | 3935 Granules

New processing version is also more likely to be up to date



MLS/Aura Level 2 Ozone (O3) Mixing Ratio V004 (ML2O3) at GES DISC

ML2O3 v004 - NASA/GSFC/SED/ESD/GCDC/GESDISC

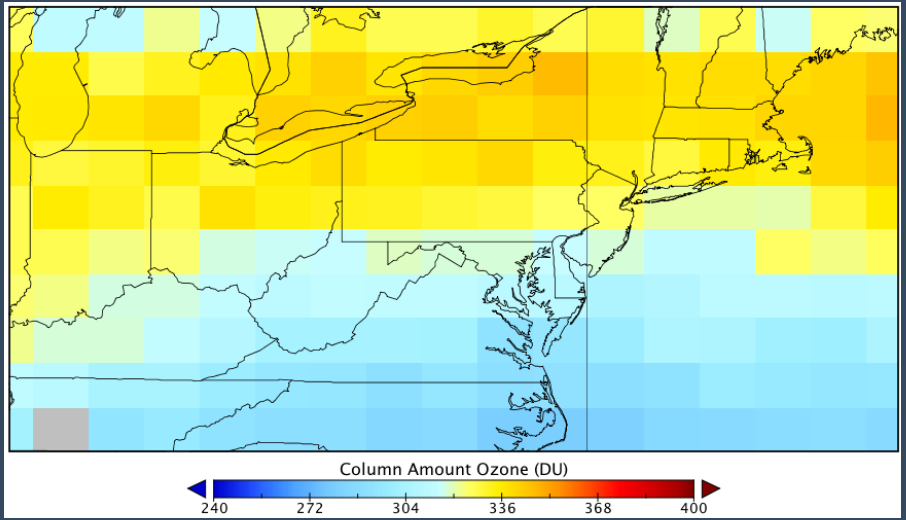
2004-08-08 ongoing | 4280 Granules

MLS/Aura Level 2 Ozone (O3) Mixing Ratio V003 (ML2O3) at GES DISC

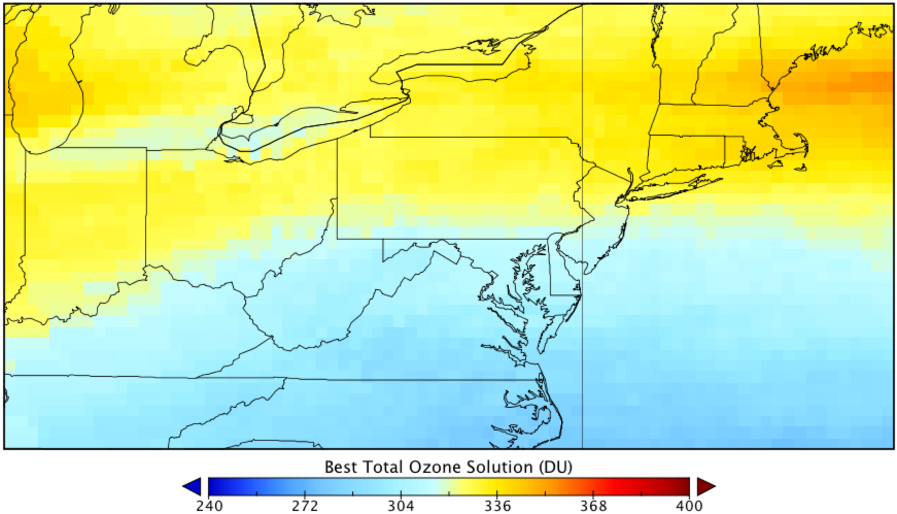
ML2O3 v003 - NASA/GSFC/SED/ESD/GCDC/GESDISC

2004-08-08 to 2015-06-30 | 3935 Granules

Newer instrument is usually better than previous instruments



Total Ozone Mapping Spectrometer



Ozone Monitoring Instrument

Region of Interest Overlap

Time Range Heuristic

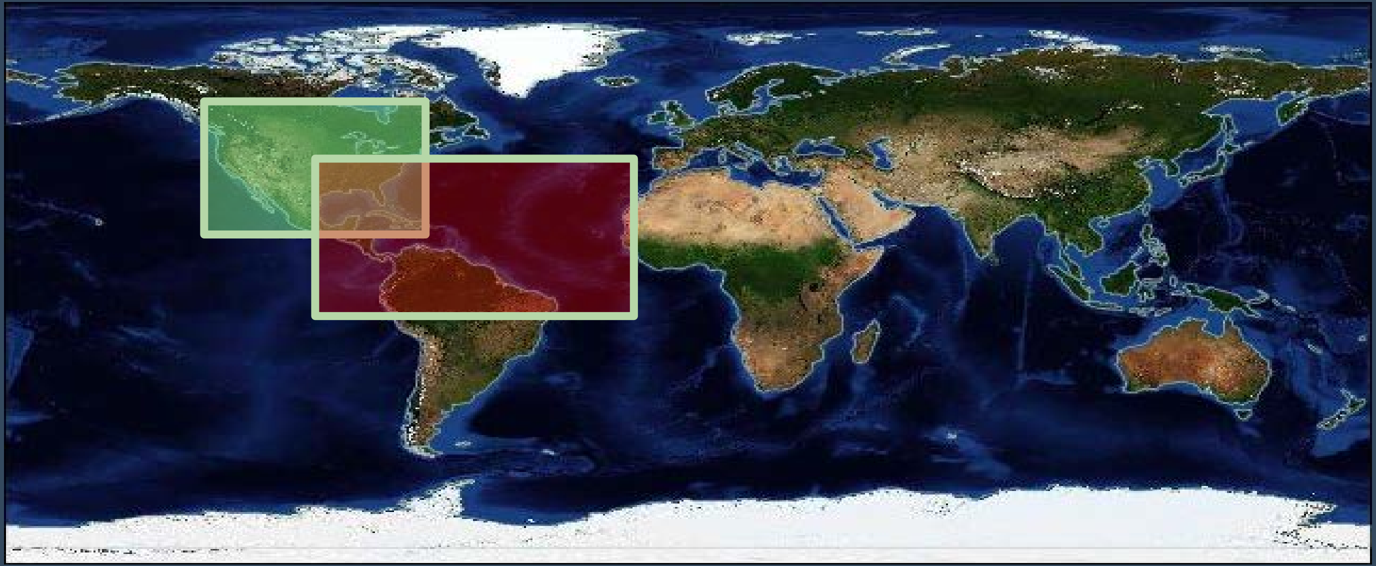
Datasets covering the user's full time range are better than those covering just part of it

		2005		2006		2007		2008		2009		2010	
Time range of interest		🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦			
TOMS-Earth Probe	🟥	🟥											Meh.
Ozone Monitoring Inst.		🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	Yeah!

Spatial Heuristic

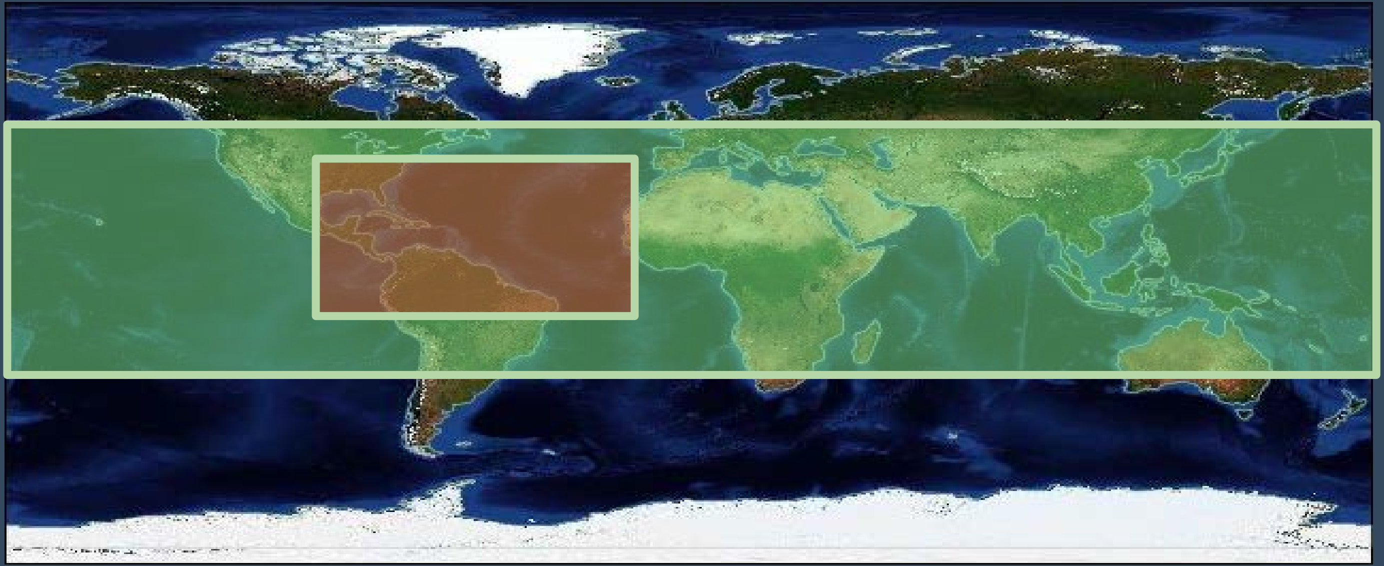
Data covering the user’s full area are better than those covering just part of it.

This is not as good as...



Spatial Heuristic

...This.



User-centric Heuristics

Community Usage Heuristic

The dataset most often used by the community is more likely to be useful

Data Product	Users**
Aqua AIRS Level 3 Daily Standard Physical Retrieval (AIRS only)*	164
Aqua AIRS Level 3 Daily Standard Physical Retrieval (AIRS+AMSU)*	714

* Version 6

** Jan 1, 2016 - June 20, 2016

User Intent Heuristics

User type or intent*	The most relevant datasets are...
Applications users	High spatial resolution, near-real-time
Students	Easier to use data *e.g., L3 grids in netCDF*
Climate Modeler	Datasets on Climate Model Grid

Digging Deeper...

Stay for the next talk, by Patrick Quinn:

“Earthdata Search: Scaling, Assessing, and Improving Relevancy”